Muzikos garso šaltinių atskyrimo giliojo mokymosi modelio SCNet apmokymas skirtingais duomenų rinkiniais

Aidas Žygas, Gražina Korvel

Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų institutas, Akademijos g. 4, Vilnius
aidas.zygas@mif.stud.vu.lt

Santrauka. Šiame straipsnyje nagrinėjama, kaip skirtingi duomenų rinkiniai (Musdb18-HQ, Moisesdb), naudojami muzikos garso šaltinių atskyrimo modelio apmokymui, lemia SCNet modelio išvesties rezultatus. Modelio efektyvumas vertinamas SDR kiekybine metrika, pagal skirtingus muzikos žanrus ir garso šaltinius. Modelis geriausiai atskiria būgnus ir vokalus, jį apmokius su Musdb18-HQ, o naudojant Moisesdb, geriausiai atskiria bosus ir būgnus. Kai modelio apmokymui ir įvertinimui naudojami skirtingi duomenų rinkiniai, SDR rodikliai tampa mažesni, nes modelis juos prasčiau apibendrina. Pastebėta, kad modeliui yra sudėtinga atskirti roko muzikos žanrą, o geriausi SDR pasiekti bliuzo ir kantri žanruose.

Raktiniai žodžiai: Muzikos garso šaltinių atskyrimas, SDR metrika, spektrograma.

1 Įvadas

Garso šaltinių atskyrimas – procesas, kurio metu iš turimo garso signalo atskiriami jį sudarantys komponentai. Garso šaltinių atskyrimas yra aktualus muzikos srityje ir yra taikomas vokalo bei instrumentų atskyrimui muzikos takelyje [1]. Norint įgyvendinti garso šaltinių atskyrimą, pasitelkiami giliojo mokymosi modeliai, kurie skirstomi į tris pagrindinius tipus, atsižvelgiant į juose įgyvendintą neuroninį tinklą [2] – konvoliucinį, rekurentinį, dėmesiu pagrįstą transformerį.

Konvoliucinį neuroninį tinklą taikantys modeliai sunkiai atpažįsta ir apdoroja ilgalaikes priklausomybes laiko srityje [5], o dėmesiu pagrįsti transformeriai yra kompleksiški ir reikalauja didelių skaičiavimų resursų [6]. Rekurentinis neuroninis tinklas modeliuoja duomenis sekomis. Šis neuroninio tinklo veikimo principas yra tinkamas garso signalų apdorojimui, nes

DOI: https://doi.org/10.15388/LMITT.2025.34

skaitmenizuotas garso signalas irgi yra reikšmių seka, išreikšta laiko srityje. Būtent šiame darbe pasirinktas SCNet [6] giliojo mokymo modelis, savo architektūroje pritaikantis rekurentinį neuroninį tinklą.

Giliojo mokymosi modelis yra apmokomas muzikos įrašų duomenimis, kurie yra pateikiami skirtingais garso takeliais. Dažniausiai apmokymui ir kokybės įvertinimui naudojamo duomenų rinkinio Musdb18-HQ vienas garso įrašas turi 5 takelius: mišrų, kuriame yra visi muzikos instrumentai ir atlikėjų vokalai, būgnų, bosų, vokalų ir likusių garso šaltinių (visų, kurie lieka iš mišraus muzikos garso signalo atmetus būgnus, bosus ir vokalus).

Svarbu modelį apmokyti ir įvertinti naudojant skirtingus duomenų rinkinius – tai padeda įvertinti modelio galimybę apibendrinti (angl. *generalize*) skirtingus muzikos žanrus ir pamatyti, kuriuose muzikos žanruose modelis negeba gerai atskirti garso šaltinių.

Šio tyrimo tikslas yra nustatyti kaip skirtingi duomenų rinkiniai, naudojami muzikos garso šaltinių atskyrimo modelio apmokymui, lemia modelio išvesties rezultatus.

Tyrimo idėja – pasirinktą SCNet giliojo mokymosi muzikos garso šaltinių atskyrimo modelį apmokyti ir įvertinti dvejais duomenų rinkiniais – Musdb18-HQ ir Moisesdb bei modelio rezultatus tarp jų – apmokyti su vienu duomenų rinkiniu, o įvertinti su kitu. Rezultatuose pateikiama modelio SDR [7] kiekybinė metrika ir nagrinėjama kaip šios metrikos reikšmė kinta esant skirtingiems muzikos žanrams ir garso šaltiniams.

2 Tyrimui pasirinktas giliojo mokymosi modelis

Tyrimui atlikti buvo pasirinktas SCNet muzikos garso šaltinių atskyrimo modelis, susidedantis iš dviejų kelių rekurentinio neuroninio tinklo [6]. Modelis susideda iš trijų pagrindinių komponentų – garso koduotojo, atskyrimo tinklo, garso dekoduotojo.

Pirmiausia garso įrašas konvertuojamas į spektrogramą, pritaikant trumpalaikę diskrečiąją Furjė transformaciją. Spektrogramą apdoroja garso koduotojas, kurio tikslas yra atskirti 3 dažnių juostas – žemų, vidutinių ir aukštų, sumažinti rezoliuciją aukštuose ir vidutiniuose dažniuose, išsaugant pilną informaciją žemuose dažniuose bei surinkti garso signalo požymius tolimesniam apdorojimui. Atskyrimo tinklas sudarytas iš dviejų kelių rekurentinio neuroninio tinklo – toliau DKRNT, architektūros, kurią modelio autoriai pritaikė remdamiesi „Dual-path RNN“ modelio architektūra [4].

DKRNT suskirsto rekurentinio neuroninio tinklo sluoksnius į du vienas kitą papildančius elementus:

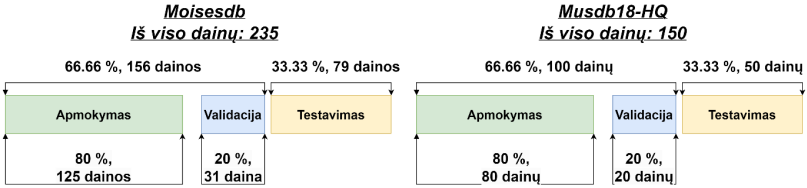
- Lokalų rekurentinį neuroninį tinklą, apdorojantį individualius segmentus atskirai, kad užfiksuoti lokalias garso savybes ir dinamiką.
- Globalų rekurentinį neuroninį tinklą, agreguojantį ir modeliuojantį informaciją dalimis, kad užfiksuoti bendrą muzikos kontekstą ir ilgalaikes priklausomybes.

Sujungimuose tarp koduotojo ir dekoduotojo pritaikomas suliejimo sluoksnis, kuris integruoja hierarchinius požymius sudėdamas dvi įvestis ir gautas rezultatas pakartojamas visoje požymių dimensijoje. Per keletą suliejimo sluoksnių atkuriami atskirti 4 garso šaltinių takeliai.

3 Duomenų rinkiniai ir jų paruošimas tyrimui

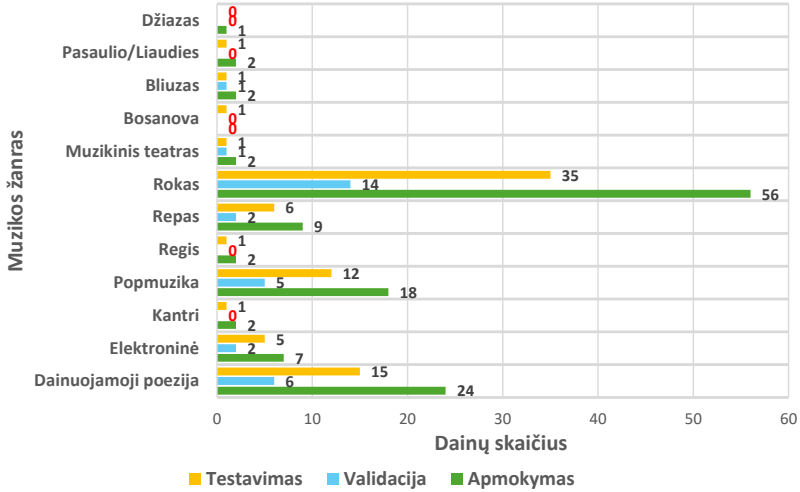
SCNet modeliui apmokyti buvo pasirinkti du duomenų rinkiniai – Musdb18-HQ [3] ir Moisesdb [1]. Abu duomenų rinkiniai susideda iš muzikos garso įrašų, įrašytų 44,1 kHz diskretizavimo dažniu, 16 bitų kvantavimo lygiu ir pateikiamų .wav failo formatu. Musdb18-HQ duomenų rinkinys susideda iš 150 pilnos trukmės skirtingų muzikos žanrų garso įrašų, iš viso sudarančių apie 10 valandų trukmę. Musdb18-HQ turi 5 garso takelius: mišrus, būgnų, bosų, vokalų ir likusių garso šaltinių [3]. Moisesdb duomenų rinkinys susideda iš 240 pilnos trukmės skirtingų muzikos žanrų garso įrašų, iš viso sudarančių apie 14 valandų trukmę [1]. Moisesdb kiekvienas garso įrašas turi tiek garso takelių, kiek skirtingų instrumentų buvo naudota. Taip pat pateikiami ir klasikiniai 5 garso takeliai kaip ir Musdb18-HQ: mišrus, būgnų, bosų, vokalų ir likusių garso šaltinių. 5 muzikos garso įrašai neturėjo bent vieno iš šių 5 garso takelių, todėl tyrimo metu buvo išimti iš duomenų rinkinio.

Modelio apmokymui, kiekvienas duomenų rinkinys buvo paskirstytas į 3 poaibius (žr. 1 pav.) – apmokymo ir validacijos, kurį sudaro 66 % viso duomenų rinkinio ir testavimo, kurį sudaro 33 % duomenų rinkinio. Apmokymo ir validacijos poaibiai tarpusavyje pasiskirstę atitinkamai 80 % ir 20 % santykiu. Kiekvieną muzikos žanrą poaibiuose siekiama paskirstyti tokiu pat santykiu kaip ir poaibiai. Dėl skirtingo dainų skaičiaus kiekviename žanre, ne visada gaunamos vienodos proporcijos. Kai tam tikro muzikos žanro dainų yra tik po vieną (pavyzdžiui džiazo ir bosanovos žanro dainos Moisesdb duomenų rinkinyje žr. 2 pav.), atsitiktiniu būdu parenkama, ar šio žanro daina bus priskirta į apmokymo ar testavimo poaibį.



1 pav. Duomenų rinkinių apmokymo validacijos ir testavimo pasiskirstymas pagal dainų skaičių

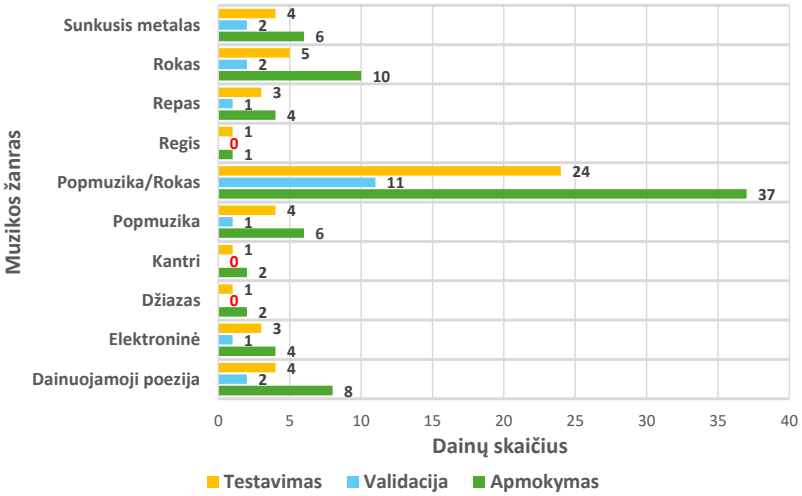
Apžvelgus Moisesdb dainų pasiskirstymą pagal žanrus (žr. 2 pav.), matoma, kad daugiausia duomenų turima roko muzikos žanro. Apmokymo validacijos ir testavimo poaibiuose yra atitinkamai 56, 14, 35 dainos. Moisesdb susideda iš vos 1 džiazo ir bosanovos žanro dainų. Tyrimo metu džiazo daina buvo priskirta apmokymo, o bosanovos – testavimo poaibiams. Dainų skaičius tarp žanrų yra nevienodas – roko žanras turi žymiai daugiau dainų, nei visi kiti.



2 pav. Dainų skaičiaus pasiskirstymas Moisesdb duomenų rinkinyje pagal apmokymo, validacijos ir testavimo poaibius.

Apžvelgus Musdb18-HQ dainų pasiskirstymą pagal žanrus (žr. 3 pav.), matoma, kad daugiausia duomenų turima popmuzikos/roko muzikos žan-

ro. Apmokymo validacijos ir testavimo poaibiuose yra atitinkamai 37, 11, 24 dainos. Musdb18-HQ turi vos 2 regio žanro dainas. Tyrimo metu regio žanro dainos buvo paskirstytos po 1 dainą į apmokymo ir testavimo poaibius. Dainų skaičius tarp žanrų yra nevienodas – popmuzikos/roko žanras turi žymiai daugiau dainų, nei visi kiti, tačiau tarp likusių žanrų pasiskirstymas yra tolygesnis, nei Moisesdb.



3 pav. Dainų skaičiaus pasiskirstymas Musdb18-HQ duomenų rinkinyje pagal apmokymo, validacijos ir testavimo poaibius.

4 Rezultatai

Modeliui apmokyti su Musdb18-HQ skirta 100 epochų, Moisesdb – 65 epochos. Abiem duomenų rinkiniams mokymosi greitis nustatytas 5e-4, partijos dydis 32. Modelio ir neuroninio tinklo konfigūracijos parametrai nebuvo optimizuojami ir keičiami. Apmokius modelį su vienu duomenų rinkiniu, buvo įvertinta jo SDR metrika (toliau – SDR) pagal skirtingus garso šaltinius ir visų garso šaltinių SDR pagal skirtingus muzikos žanrus. SDR parodo santykį tarp atskaitos signalo ir garso šaltinių atskyrimo metu susidariusių triukšmų ir išreiškiama decibelais (dB) [7]. Įvertinimui naudojamas to paties duomenų rinkinio testavimo poaibis ir nekeičiant apmokymo duomenų rinkinio naudojamas kito duomenų rinkinio testavimo poaibis. Žemiau pateikiami gauti rezultatai (žr. 1 lentelę, žr. 2 lentelę).

1 lentelė. Visų žanrų vidutinis garso šaltinių SDR pagal skirtingus duomenų rinkinių poaibius.

Apmokymo (AP) ir įvertinimo (ĮV) duomenų rinkinys	Bosai	Būgnai	Kiti garso šaltiniai	Vokalai	Visi garso šaltiniai
AP: Mudb18-HQ, ĮV: Musdb18-HQ	8,90	9,90	6,46	9,74	8,75
AP: Musdb18-HQ, ĮV: Moisesdb	10,87	10,93	7,07	10,17	9,76
AP: Moisesdb, ĮV: Moisesdb	**11,99**	**11,90**	**7,88**	**10,74**	**10,63**
AP: Moisesdb, IV: Musdb18-HQ	8,31	9,27	5,88	9,70	8,29

2 lentelė. Skirtingų duomenų rinkinių, naudotų apmokymui ir įvertinimui visų garso šaltinių vidutinė SDR pagal žanrus.

Žanras	AP: Mudb18-HQ, ĮV: Musdb18-HQ	AP: Musdb18-HQ, ĮV: Moisesdb	AP: Moisesdb, ĮV: Moisesdb	AP: Moisesdb, ĮV: Musdb18-HQ
Bliuzas	-	**13,14**	**14,57**	-
Elektroninė	6,18	7,22	8,10	6,14
Kantri	**12,85**	9,97	10,79	**13,57**
Rokas	8,71	9,09	9,71	8,92

Naudojant Musdb18-HQ duomenų rinkinį apmokymui/testavimui, geriausiai išskirti garso šaltiniai yra būgnai ir vokalai, kurių SDR yra 9,9 dB ir 9,74 dB atitinkamai. Naudojant Moisesdb duomenų rinkinį apmokymui/testavimui, geriausiai išskirti garso šaltiniai yra bosai ir būgnai, kurių SDR yra 11,99 dB ir 11,90 dB atitinkamai. Visais duomenų rinkinių poaibių atvejais, modelis prasčiausiai išskiria kitų garso šaltinių garso takelį ir dažniausiai prasčiausiai išskiria elektroninės muzikos žanrą. Turint skirtingus apmokymo ir testavimo poaibius, naudojant Musdb18-HQ rinkinį apmokymui, o Moisesdb testavimui, gaunama aukštesnė SDR, nei apmokymui naudojant Moisesdb ir testavimui Musdb18-HQ. Kai įvertinimui naudojamas Moisesdb duomenų rinkinio testavimo poaibis, geriausiai išskiriami bliuzo žanro garso šaltiniai, nors bliuzo apmokymo ir validacijos poaibiuose buvo vos po vieną dainą, o testavimo – 2 dainos. Kai apmokymui naudojamas Moisesdb duomenų rinkinys, o testavimui Musdb18-HQ, geriau atskiriami kantri muzikos žanro garso šaltiniai, nei apmokymui naudojant Musdb18-HQ, nors

visų žanrų skirtingų garso šaltinių vidutinė SDR visada žemesnė, naudojant skirtingus apmokymo ir testavimo duomenų rinkinius. Daugiausia abiejuose duomenų rinkiniuose apmokymo poaibiuose yra roko muzikos žanro dainų, tačiau SDR šiame žanre niekada nėra aukščiausia.

5 Išvados

SCNet modelį apmokius pasirinktais dviem duomenų rinkiniais ir įvertinus modelio garso šaltinių atskyrimo muzikoje rezultatus, gautos šios išvados:

- SCNet modelis negali gerai apibendrinti skirtingų apmokymo ir testavimo duomenų rinkinių, nes SDR visada būna žemesnė, nei modelį apmokant ir testuojant su tuo pačiu duomenų rinkiniu. Modelio, apmokyto ir testuoto su Musdb18-HQ duomenų rinkinio poaibiais visų garso šaltinių SDR yra 8,75 dB. Tuo tarpu naudojant Musdb18-HQ testavimo poaibį ir modelį apmokius su Moisesdb apmokymo poaibiu, SDR sumažėja iki 8,29 dB. Panaši tendencija matoma ir naudojant Moisesdb testavimo rinkinį. Modelį apmokius su Moisesdb apmokymo poaibiu, visų garso šaltinių SDR yra 10,63 dB, o su Musdb18-HQ SDR yra 9,76 dB.
- Roko muzikos žanras yra sudėtingas jį sudarančių garso šaltinių atžvilgiu ir didelis duomenų kiekis (56 dainos Moisesdb ir 10 dainų Musdb18-HQ apmokymo poaibiuose) nelemia geriausių garso šaltinių atskyrimo rezultatų tarp skirtingų muzikos žanrų. Roko muzikos žanro visų garso šaltinių SDR yra nuo 8,71 dB iki 9,71 dB, priklausomai nuo skirtingų apmokymo ir testavimo duomenų rinkinių poaibių.
- Muzikos žanrai, turintys nedidelį apmokymo duomenų kiekį kaip bliuzas ir kantri, pasiekia aukštą SDR. Bliuzo SDR yra 14,57 dB, naudojant Moisesdb apmokymo ir testavimo rinkinių poaibius, o kantri SDR yra 13,57 dB, naudojant Moisesdb apmokymo ir Musdb18-HQ testavimo poaibius. Tai rodo, kad didelis apmokymo duomenų kiekis nebūtinai užtikrina geresnius garso šaltinių atskyrimo rezultatus ir signalizuoja, kad panašūs bliuzo ir kantri muzikos žanrai tarpusavyje turi artimas akustines savybes dėl kurių modelis gali pasiekti aukštesnę visų garso šaltinių SDR.
- Moisesdb duomenų rinkiniui reikėjo mažiau epochų (65), nei Musdb18-HQ (100), kad pasiekti panašius SDR rezultatus modelį įvertinant su Musdb18-HQ testavimo poaibiu. Moisesdb duomenų rinkinys yra di-

desnis ir susideda iš 235, o Musdb18-HQ iš 150 dainų. Tai rodo, kad Moisesdb duomenų rinkinio didesnė duomenų įvairovė lemia modelio greitesnį mokymąsi per epochas. Taip pat, didesnis modelio apmokymo epochų skaičius ir mažesnis duomenų kiekis lemia geresnį modelio apibendrinimą, kadangi Musdb18-HQ visų garso šaltinių SDR buvo aukštesnė, kai testavimo poaibis buvo pakeistas iš Musdb18-HQ į Moisesdb (9,76 dB > 8,75 dB). Moisesdb atveju buvo atvirkščiai, SDR buvo žemesnė (8,29 dB < 10,63 dB).

Padėka. Dėkojame Vilniaus universiteto ITAPC padaliniui už suteiktus IT išteklius (HPC), kurie leido greičiau apmokyti modelius bei supaprastino viso tyrimo atlikimą. Tyrimas finansuojamas pagal LR Švietimo, mokslo ir sporto ministerijos programą „Universitetų ekscelencijos iniciatyvos“ (LR ŠMSM mokslo plėtros programos pažangos priemonė Nr. 12-001-01-01-01 „Gerinti mokslo ir studijų aplinką“).

Literatūra

[1] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, “Moisesdb: A dataset for source separation beyond 4-stems”, arXiv.org, 2023, doi: https://doi.org/10.48550/arXiv.2307.15913.

[2] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, “Music Source Separation with Band-Split RoPE Transformer”, arXiv.org, 2023, doi: https://doi.org/10.48550/arXiv.2309.02612.

[3] Z. Rafii, A. Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and R. Bittner, “MUSDB18 - a corpus for music separation”, Zenodo (CERN European Organization for Nuclear Research), 2017, doi: https://doi.org/10.5281/zenodo.1117372.

[4] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation”, arXiv.org, 2019, doi: https://doi.org/10.48550/arXiv.1910.06379.

[5] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, “Audio Captioning Transformer,” arXiv.org, 2021. https://arxiv.org/abs/2107.09817.

[6] W. Tong et al., “SCNet: Sparse Compression Network for Music Source Separation”, arXiv.org, 2024, doi: https://doi.org/10.48550/arXiv.2401.13276.

[7] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation”, IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 4, pp. 1462–1469, Jul. 2006, doi: https://doi.org/10.1109/tsa.2005.858005.